

ANTONIO ŠIMIĆ

AI Engineer

Zagreb, Croatia (open to remote / relocation) | +385 97 714 0610 | antonio.simic@fer.hr | antoniosimic.site
[linkedin/antoniosimic](https://www.linkedin.com/in/antoniosimic) | github.com/antoniosimic

SUMMARY

AI Engineer with hands-on experience shipping production LLM agents and RAG systems on Google Cloud, currently building a conversational AI agent for Croatia's largest retail chain, architected for 150,000+ users. Comfortable across the full stack — model/tool design, multi-agent orchestration, cloud deployment, MCP servers, and production debugging. 1st-place hackathon winner in computer vision; founder of an ed-tech platform serving 800+ students.

WORK EXPERIENCE

AI Engineer / Data Scientist @ mStart - Fortenova Group	2025 – Present
• Multi-Agent Architecture: Designed a multi-agent system (root agent + specialized shopping sub-agent) on Google ADK/Gemini, with controlled hand-off logic and loop-prevention guards for reliable multi-agent conversations. • Production AI Agent for Konzum: Built a conversational AI agent on WhatsApp for Konzum, Croatia's largest retail chain, delivering real-time personalized shopping recommendations and store locations — architected to scale to 150,000+ users. • RAG, Live Data & Personalization: Implemented a Hybrid Vector Search RAG system on AlloyDB with an automated daily refresh pipeline and validation checks; integrated Vertex AI Memory Bank for persistent cross-session personalization (dietary preferences, shopping habits, saved store); tuned AlloyDB/PostgreSQL/LLMs, cutting average response latency ~30% (10.5s → 7s). • Safety & Security: Built a two-stage safety classifier (keyword pre-filter + cached LLM classification) screening messages for prompt injection, abuse, and scope violations; implemented authentication, rate limiting, and guardrails. • Reliability & Debugging: Diagnosed and fixed critical production bugs (hallucinated basket totals, cache invalidation, Agent Engine quota-driven 429 errors); built context/event compaction to keep long-running sessions within token limits. • Infrastructure & CI/CD: Built the WhatsApp integration (webhook + Pub/Sub worker with HMAC verification) and a multi-environment CI/CD pipeline across dev/test/prod GCP projects; engineered a custom MCP server and internal web app to interface with the Agent Engine. • Observability, Evals & Cost: Set up logging, dashboards, and alerting; built an evaluation system (regression tests, A/B testing) for response quality; optimized LLM cost through caching, model routing, and batching. • AI-Native Development: Used Claude Code with custom subagents and skills to accelerate development velocity.	
Founder & Lead Developer, Prepko	2022 – 2026
• Architected and developed a full-stack online learning platform from the ground up to teach and mentor first-year FER students. • Scaled the platform to over 800 users. • Engineered a custom WordPress CMS with themes and plugins using PHP, JavaScript, and MySQL to deliver a scalable, high-quality educational platform.	

PROJECTS

Multi-Agent Assistive System for the Visually Impaired FER, Software Agents course	2026
• Built a multi-agent system combining GPT-4o Vision (visual analysis), Claude Opus (semantic interpretation), and speech synthesis, orchestrated with LangGraph and ChromaDB vector memory. • Full-stack implementation with a Next.js frontend and FastAPI backend.	
Road Network Extraction 1st Place, Neodata Hackathon	2024
• Developed a computer vision model to extract road networks from a critically small dataset of only 50 satellite images. • Designed a data augmentation strategy (rotations, mirroring, skewing) to expand the dataset 32x. • Built and optimized an end-to-end image processing pipeline under a strict 24-hour time constraint.	

EDUCATION

Master of Science in Data Science Faculty of Electrical Engineering and Computing (FER)	2025 – 2027
Bachelor in Computer Engineering Faculty of Electrical Engineering and Computing (FER)	2022 – 2025

TECHNICAL SKILLS

- **AI & Machine Learning:** LLMs, Agentic Frameworks (LangGraph, LangChain, Google ADK, A2A protocol), RAG, Hybrid Embeddings & Vector Search, PyTorch, TensorFlow, Computer Vision, Reinforcement Learning.
- **LLM Engineering & Ops:** Prompt Optimization, Evals & Regression Testing, Latency & Cost Optimization, Observability, Guardrails/Prompt-Injection Protection, Claude Code (custom subagents & skills).
- **Cloud & DevOps:** Google Cloud Platform (GCP), Cloud Run, Vertex AI Agent Engine, Git, Docker, Kubernetes.
- **Databases:** AlloyDB, ChromaDB, Vector Databases, PostgreSQL, MySQL.
- **Web Development & Tools:** MCP, React, Next.js, Flask, Spring, REST/WhatsApp APIs.